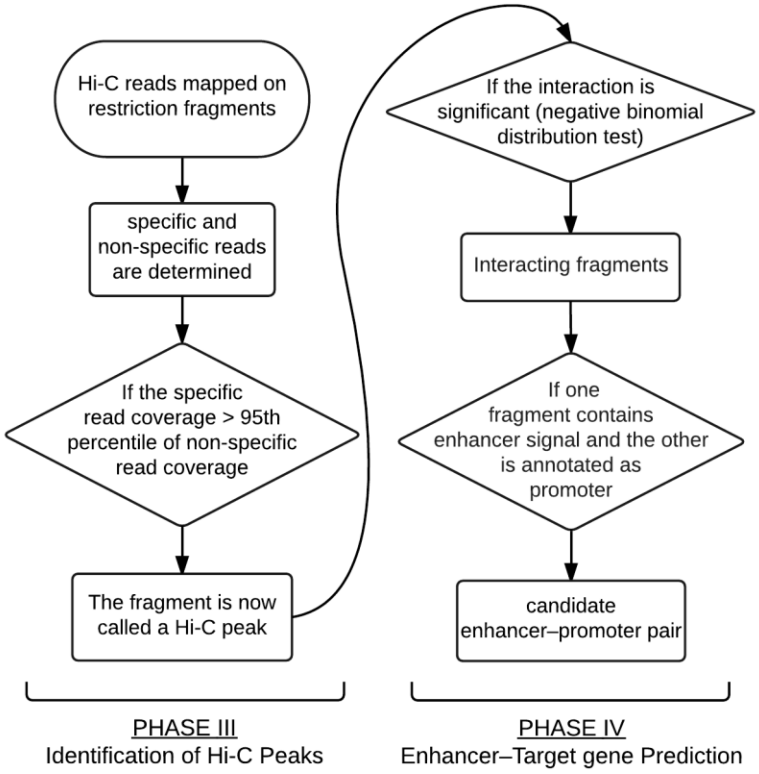
List of Supplementary Information

---

0. Supplementary Figure



Supplementary Figure 1: The quality control flow for HIPPIE phase III and phase IV.

1. Coverage threshold for restriction fragments (Hi-C peak identification)

We first identified specific and non-specific read pairs by the distances of each mapped read in a pair from the closest restriction site (d_1 and d_2). When $d_1 + d_2 \leq 500$ nt, both reads are considered specific reads, while $d_1 + d_2 > 500$ nt, both reads are considered non-specific reads as previously described (Yaffe & Tanay, 2011). We then calculated the specific and non-specific read coverage of each restriction fragment. A Hi-C peak is called if the specific read coverage for a restriction fragment is higher than the 95th

percentile of the non-specific read coverage distribution. Non-specific reads are subsequently discarded and only Hi-C peaks and the specific reads denoting their interaction partners are used for further analysis.

2. Calculating mappability and GC content

We generated a set of 36 base pair (bp) pseudo-reads using a one nucleotide sliding window to extract reads from both ends (500 nt in length) of each restriction fragment (i.e. the region between two consecutive restriction enzyme cut sites along the human genome). The pseudo-reads are then mapped back to the genome using BWA and the mappability of each restriction fragment is determined by the fraction of uniquely mapped pseudo-reads (MAPQ ≥ 30 and with SAM tag of "XT:A:U" in BWA) for both of its ends. The GC content of each fragment is calculated as the percentage of total bases that are either guanine or cytosine within the 500 nts that make up their ends (Jin et al., 2013).

3. Estimating random contact frequencies and significance calculation

We implemented the calculation of random contact frequencies described previously (Jin et al., 2013) into our HIPPIE analysis pipeline. Specifically, the expected number of read pairs ($\mu_{i,j}$) for each interaction (*i, j*) on the same chromosome is:

$$\mu_{i,j} = m_i \times m_j \times F_{i,j}^{gc} \times L_{i,j}$$

We set $x_{i,j}$ as the observed (actual) read pair supporting the interaction of restriction fragments *i* and *j*. In order to account for the inherent biases of the Hi-C methodology, we first binned the restriction fragments by the GC content of their ends into 20 bins (with break points 0, 0.05, 0.01,…, 0.90, 0.95, and 1). For the length of the restriction fragments, we took the log of this value and binned by 2 orders of magnitude. If two restriction fragments are on the same chromosome, we binned the distance of each by a 5000 bp window size. We then let B_i^{len} be the bin assignment of restriction fragment *i* by its length, B_i^{gc} to allow the bin assignment of restriction fragment *i* by its GC content, and $B_{i,j}^{gc}$ be the bin assignment for the fragment interaction of *i* and *j* based on their linear distance.

Then $L_{i,j}$ is the expected frequency of contacts between fragment *i* and *j*, using a correction factor for restriction fragment length bias:

$$L_{i,j} = \frac{\sum_{k,l} \frac{x_{k,l}}{m_k \times m_l}}{\sum_{k,l} 1}$$

Where

$\forall\{k,l\}$ satisfy: $B_k^{len} = B_i^{len}, B_l^{len} = B_j^{len}, B_{k,l}^{dist} = B_{i,j}^{dist}, chr(k) = chr(l), m_k > 0.2, \text{and } m_l > 0.2$

Where $F_{i,j}^{gc}$ is a correction factor for GC content bias (contact fraction for the corresponding GC bin among all possible GC bins) between fragment *i* and *j* :

$$F_{i,j}^{gc} = \frac{\sum_{k,l} \frac{x_{k,l}}{m_k \times m_l} / \sum_{k,l} 1}{\sum_{u,v} \frac{x_{u,v}}{m_u \times m_v} / \sum_{u,v} 1}$$

Where $\forall\{k,l\}$ satisfy: $B_k^{gc} = B_i^{gc}, B_l^{gc} = B_j^{gc}, B_{k,l}^{dist} = B_{i,j}^{dist}, chr(k) = chr(l), m_k > 0.2, \text{and } m_l > 0.2$, and $\forall\{u,v\}$ satisfy: $chr(u) = chr(v), m_u > 0.2, \text{and } m_v > 0.2.$

Note for inter-chromosomal interactions, the same estimation equations are used as above, except the requirements of $chr(k) = chr(l),\ chr(u) = chr(v), \text{and } B_{k,l}^{dist} = B_{i,j}^{dist}$.

With the estimation of $\mu_{i,j}$ of each restriction fragment pair (*i, j*), we then fit all $X_{i,j}$ to a negative binomial distribution to estimate the statistical significance of the interaction between each pair:

$$X_{i,j} \sim NB(u_{i,j}, p)$$

Where *p* is the fixed value (*β*-1/*β*), where *β*=2.057 as derived by (Jin et al., 2013)).

4. Supplementary references

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., … Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, *503*(7475), 290–4. doi:10.1038/nature12644

Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*. doi:10.1038/ng.947